Mechanically separated HAM and SPAM

April 27, 2015

Overview and Motivation

- Project: Use machine learning to predict whether emails are SPAM or not ("HAM")
- Data: a subset of the Enron email corpus and a curated set of SPAM email (with malicious aspects removed)
    - http://www.aueb.gr/users/ion/data/enron-spam/
- Motivation: want to learn how to "quantify" text in order to apply machine learning techniques to extract information from large amounts of documents

Work plan

- Preprocess data: have raw emails; need to preprocess in such a way that features (words and phrases) can be easily extracted and accurately

- Feature selection: convert text files to word vectors

    - my current code produces a huge number of potential features (n=33,801 , p=157,311), many of which are probably useless or duplicative

    - part of this process will be determining how to reduce the set of possible features to the most useful set for making predictions

- Model selection: find the model that is able to best predict whether an email is SPAM

Current issues

- What is unicode and why is it breaking my code?
- So much metadata (To, From, Encoding, …)
- Need to extract words and phrases more consistently and more efficiently
    - the word “don’t” could show up in the feature set as “do”, “not”, “dont”, “don”, “t”, etc.
    - intend to experiment with ‘stemming’ to reduce size of feature set
    - are numbers important? or should they be dropped?

Advice and comments always appreciated

https://github.com/jw-ml/dat5_spam-filter